Groupwise Segmentation with Multi-atlas Joint Label Fusion

Hongzhi Wang and Paul A. Yushkevich*

Department of Radiology, University of Pennsylvania

Abstract. Groupwise segmentation that simultaneously segments a set of images and ensures that the segmentations for the same structure of interest from different images are consistent usually can achieve better performance than segmenting each image independently. Our main contribution is that we adopt the groupwise segmentation framework to improve the performance of multi-atlas label fusion. We develop a novel statistical model to allow this extension. Comparing to previous atlas propagation and groupwise segmentation work, one key novelty of our method is that the error produced during label propagation is explicitly addressed in the joint label fusion framework. Experiments on hippocampus segmentation in magnetic resonance images show the effectiveness of the new groupwise segmentation technique.

1 Introduction

As a primary mechanism for quantifying the properties of anatomical structures and pathological formations from imaging data, image segmentation is an important task in medical image analysis. Typically, a segmentation algorithm is applied to segment one image at a time, i.e. segmenting one image is independent from segmenting other images. However, different segmentation tasks may not be independent, especially when images share common structures and similar appearances. When some images share similarities, one may expect that their segmentations should be related as well. By enforcing consistency in the segmentations produced for them, one may improve the robustness of automatic segmentation against random effects.

The idea of incorporating region coherence of same or similar objects across different images to reduce segmentation errors was initially addressed in the joint registration and segmentation framework, e.g. [2,7,10], motivated by the observation that image registration and image segmentation are highly correlated tasks. Improving one can help improve the other. By registering multiple images into a common space, appearance models of the same structure of interest from all images can be collected and re-enforced to ensure similar appearances for the segmented structures from different images. The estimated segmentations can then be used to improve registration such that segmentation alignments

* This work was supported by NIH awards AG037376, EB014346.

K. Mori et al. (Eds.): MICCAI 2013, Part I, LNCS 8149, pp. 711–718, 2013.

after registration are improved. With registrations between testing images, both appearance and shape consistencies can be enforced in groupwise segmentation.

Since the groupwise segmentation idea can be implemented with any non-groupwise segmentation algorithms, to directly improve upon the state of the art medical image segmentation techniques, we adopt the groupwise segmentation framework to improve the performance of multi-atlas label fusion (MALF). Through establishing one-to-one correspondence between a target image and a pre-labeled image, i.e. *atlas*, by image-based deformable registration, MALF transfers segmentation labels from the atlas to the target image and uses label fusion to combine solutions produced by different atlases. The highly competitive performance over many challenging applications, e.g. [5,13,15], show that the example-based knowledge representation and registration-based knowledge transfer model employed by MALF can produce highly accurate segmentation for medical applications.

Since pairwise or groupwise registration among testing images are often required in groupwise segmentation, it is a natural extension to apply techniques developed in MALF for groupwise segmentation. Our first contribution is to realize this extension through a novel statistical model for groupwise segmentation. Similar to the atlas propagation work [15] and the recent groupwise segmentation work [6], in our approach, each testing image and its estimated segmentation is applied as an additional atlas to help improve the segmentation accuracy for other testing images. However, when a testing image is applied as an atlas, due to the errors in producing its automatic segmentation, it is expected to be less reliable than the original atlases. Our second contribution is to extend the joint label fusion technique [14] to address this limitation. For validation, we apply our approach to segment the hippocampus from MRI and show significant performance improvements over MALF and other label propagation work.

2 Methods

Image segmentation can be addressed via estimating the conditional probability $p(T_S|T_F, \mathcal{D})$, where T_F is the image to be segmented, T_S is a segmentation for T_F and $\mathcal{D}$ is the training data, which, for example, may include some images and their gold standard segmentations. The conditional probability can be estimated through various methods, e.g. discriminative learning or MALF.

In the MALF framework, $\mathcal{D}$ contains all atlases. The conditional probability is estimated in the form $p(l|x, T_F, \mathcal{D})$ through warping each atlas to the target image, followed by label fusion, where l indexes through all labels and x indexes through all voxels in T_F. This technique is described in detail below. With accurate conditional probabilities, the true segmentation can be estimated by $T_S(x) = \text{argmax}_l p(l|x, T_F, \mathcal{D})$.

2.1 Multi-atlas Label Fusion

Since our work is based on MALF, we briefly describe the technique. Let $\mathcal{D} = \{A^1 = (A^1_F, A^1_S), ..., A^n = (A^n_F, A^n_S)\}$ be n atlases, warped to the space of a

target image by deformable registration. A_F^i and A_S^i denote the i_{th} warped atlas image and manual segmentation.

One simple and powerful label fusion technique is weighted voting, where each atlas contributes to the final solution according to a weight. Among the weighted voting approaches, similarity-weighted voting strategies with spatially varying weight distributions have been particularly successful [1,13,14]. The consensus votes received by label l are:

$$\hat{p}(l|x, T_F, \mathcal{D}) = \sum_{i=1}^{n} w_x^i p(l|x, A^i) \tag{1}$$

where $\hat{p}(l|x, T_F, \mathcal{D})$ is the estimated probability of label l for the target image. $p(l|x, A^i)$ is the probability that A^i votes for label l at x, with $\sum_{l\in\{1,\ldots,L\}} p(l|x, A^i) = 1$. Typically, for deterministic atlases that have one unique label for every location, $p(l|x, A^i)$ is 1 if $l = A_S^i(x)$ and 0 otherwise. w_x^i is the voting weight for the i_{th} atlas, with $\sum_{i=1}^{n} w_x^i = 1$.

To estimate voting weights, similarity metrics employed by image-based registration such as sum of squared distance and normalized cross correlation can be applied such that atlases with more similar appearance to the target image at location x receives higher votes. One limitation of this approach is that the voting weights for each atlas is estimated independently from other atlases, ignoring potential correlations among the atlases. To address this problem, the joint label fusion algorithm estimates voting weights by simultaneously considering pairwise atlas correlations. As shown in [14], joint label fusion performed better than label fusion with independent weight estimation.

In the joint label fusion approach, segmentation errors produced by one atlas are modeled as $T_{S,l}(x) = A_{S,l}^i(x) + \delta^i(x)$. $T_{S,l}(x), A_{S,l}^i(x) \in \{0, 1\}$ are the observed votes for label l produced by the target image and the i_{th} warped atlas, respectively. Hence, $\delta^i(x) \in \{-1, 0, 1\}$ is the observed label error. Note that both $T_{S,l}$ and $\delta^i(x)$ are unknown. The probability that different atlases produce the same label error at location x are captured by a dependency matrix M_x, with $M_x(i, j) = p(\delta^i(x)\delta^j(x) = 1 \mid T_F, A_F^i, A_F^j)$ measuring the correlation between i_{th} and j_{th} atlases. In [14], the pairwise atlas correlation is estimated by appearance correlation as $M_x(i, j) \sim \left[\sum_{y\in\mathcal{N}(x)} |A_F^i(y) - T_F(y)||A_F^j(y) - T_F(y)|\right]^{\beta}$, where $\mathcal{N}(x)$ defines a neighborhood around x and β is a model parameter. The expected label difference between the combined solution and the target segmentation is:

$$E\left[(T_{S,l}(x) - \sum_{i=1}^{n} \mathbf{w}_x^i A_{S,l}^i(x))^2 \mid F_T, F_1, ..., F_n\right] = \mathbf{w}_x^t M_x \mathbf{w}_x \tag{2}$$

where t stands for transpose. To minimize the expected label difference, the voting weights are solved by $\mathbf{w}_x = \frac{M_x^{-1} 1_n}{1_n^t M_x^{-1} 1_n}$, where $1_n = [1; 1; ...; 1]$ is a vector of size n.

2.2 Formulation for Groupwise Segmentation

Let $\mathcal{T}_F = \{T_F^1, ..., T_F^m\}$ be m testing images to be segmented. Groupwise segmentation can be formulated as jointly segmenting all testing images using the training set $\mathcal{D}$ and can be solved via estimating the joint conditional probability $p(\mathcal{T}_S = \{T_S^1, ..., T_S^m\}|\mathcal{T}_F, \mathcal{D})$, where $T_S^1, ..., T_S^m$ are the estimated segmentations for $T_F^1, ..., T_F^m$, respectively. Since it is difficult to directly estimate the joint probability, we apply the pseudolikelihood approximation technique [3] and estimate the joint probability by: $p(\mathcal{T}_S|\mathcal{T}_F, \mathcal{D}) =$

$$\prod_{k=1}^{m} p(T_S^k|\mathcal{D}, \mathcal{T}_F, \mathcal{T}_S\backslash\{T_S^k\}) = \prod_{k=1}^{m}\prod_{x} p(l^k(x)|x, \mathcal{D}, \mathcal{T}_F, \mathcal{T}_S\backslash\{T_S^k\}) \tag{3}$$

where $l^k(x) = \text{argmax}_l p(l|x, \mathcal{D}, \mathcal{T}_F, \mathcal{T}_S\backslash\{T_S^k\})$ is the estimated label for the k_{th} testing image. In this model, we assume that the label probability for each voxel is conditionally independent given the images and segmentations in (3). Note that the segmentation of each testing image is estimated by both the original atlases and the remaining testing images. Hence, the correlations between testing images are explicitly considered to make their solutions compatible.

Like the pseudolikelihood approach, the segmentations for all testing images are computed through iterative estimation. First, the segmentation of each testing image is independently estimated with MALF only using the atlases. In each of the following iterations, the estimated segmentation for each testing image is updated one at a time to maximize the joint probability (3). Using weighted voting based label fusion, we estimate the label probability for one testing image T_F^k by:

$$p(l|x, \mathcal{D}, \mathcal{T}_F, \mathcal{T}_S\backslash\{T_S^k\}) = \sum_{i=1}^{n} \mathbf{w}^i A_{S,l}^i + \sum_{j=1, j\neq k}^{m} \mathbf{w}^{a^j} a_{S,l}^j \tag{4}$$

where a^j is the candidate segmentation produced by warping the segmentation produced for the j_{th} testing image to T_F^k. w^{a^j} is the voting weight assigned to it, with $\sum_{i=1}^{n} \mathbf{w}^i + \sum_{j=1, j\neq k}^{m} \mathbf{w}^{a^j} = 1$. Note that, for a simpler notation, the parametrization by x is implicit.

Potential risk in using testing images as atlases. Due to registration and label fusion errors, it is reasonable to expect that the segmentation produced for each testing image is less accurate than those of the original atlases. Hence, when applying a testing image as an atlas to segment other testing images, in addition to the errors produced by image-based deformable registration, segmentation errors produced for the testing image are also propagated to other testing images. This potential risk may result in overall less accurate candidate segmentations produced by warping a testing image than by directly warping an original atlas.

Image similarity based label fusion is effective for detecting and reducing segmentation errors caused by registration errors. However, it can not detect whether the atlas contains errors in its segmentation. To address the unreliability of using testing images as additional atlases, we propose a solution based on the

following observation. If an atlas produces more segmentation errors than other atlases, it is expected that its voting weight should be smaller than other atlases in the optimal solution. We propose to incorporate such prior knowledge in similarity-based weighted voting for more robust label fusion. To this end, we explicitly control the contribution from testing images in (4). Following the joint label fusion technique, we estimate the label probability for each testing image by solving the following optimization problem:

$$E\left[(T^k_{S,l} - \sum_{i=1}^{n} \mathbf{w}^i A^i_{S,l} - \sum_{j=1,j\neq k}^{m} \mathbf{w}^{a^j} a^j_{S,l})^2 \mid \mathcal{D}, \mathcal{T}_F\right] = [\mathbf{w}; \mathbf{w}^a]^t M [\mathbf{w}; \mathbf{w}^a]$$

$$\text{subject to } \sum_{i=1}^{n} \mathbf{w}^i = 1 - \lambda, \qquad \sum_{j=1,j\neq k}^{m} \mathbf{w}^{a^j} = \lambda \quad (5)$$

For segmenting one testing image, the contribution from the remaining testing images is controlled by the total weight assigned to them, $0 \leq \lambda < 1$. Typically, when the atlases cannot produce reliable segmentation, one may expect more contribution from testing images to regularize the results and vice versa.

Applying Lagrange multipliers, we can solve (5) in closed form by:

$$[\mathbf{w}; \mathbf{w}^a]^t = M^{-1}(\mu_c c + \mu_d d) \quad (6)$$

where $c = [1; ...; 1; 0; ...; 0], d = [0; ...; 0; 1; ...; 1]$. Only the first n entries in c and the last $m-1$ entries in d are non-zero. $\begin{bmatrix} \mu_c \\ \mu_d \end{bmatrix} = \begin{bmatrix} c^t M^{-1} c, \; c^t M^{-1} d \\ d^t M^{-1} c, \; d^t M^{-1} d \end{bmatrix}^{-1} [1-\lambda; \lambda]$.

3 Experiments

Imaging data and Experiment setup. Our study is conducted using 1.5 T baseline MR images from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Among these images, 57 are normal controls, 84 are patients with mild cognitive impairment (MCI) and 41 are patients with AD. Manual segmentations of the hippocampus are provided by ADNI as well. For cross-validation evaluation, we randomly selected 10 images to be the training images, i.e. the atlases, and another 50 images for testing. The cross-validation experiment was repeated for five times. In each experiment, a different set of atlases and testing images were randomly selected. The results reported below are averaged over the five experiments. To examine the performance with respect to the number of atlases used for producing the initial segmentation, in each cross-validation experiment, we also tested with different numbers of atlases, varying from 1 to 10.

For label fusion, we applied a $5\times5\times5$ neighborhood for $\mathcal{N}$. In our experiments, we fixed $\beta = 2$ for computing the atlas correlation matrix, which is shown to be optimal in [14] for hippocampus segmentation. For groupwise segmentation, we fixed $\lambda = 0.5$. Hence, the expected contribution from each testing image is significantly smaller than the contribution from each atlas.

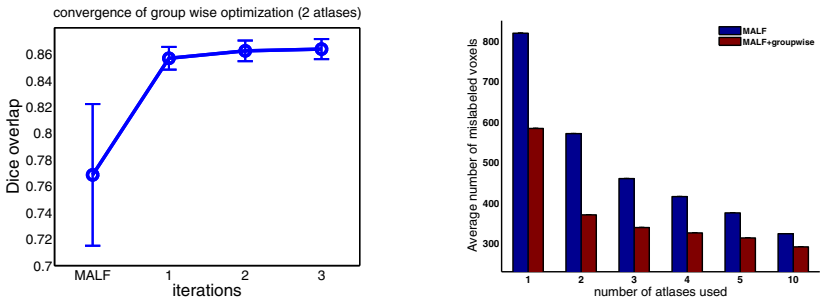

Fig. 1. Left: Segmentation accuracy (in terms of Dice similarity coefficient $\frac{2|A\cap B|}{|A|+|B|}$) of our groupwise segmentation algorithm at each iteration. The results are averaged over 5 cross-validation experiments. Error bar is at 0.25 standard deviation. The segmentation produced by MALF used two atlases; Right: Segmentation performance in terms of average number of mislabeled voxels per hippocampus.

Results. As shown in Fig. 1, our iterative optimization usually converges within a few iterations, with the first iteration producing the maximal performance improvement and dramatic diminishing performance gains in later iterations. In our experiment, we set the maximal iteration number to be 3. Fig. 1 also shows the performance produced by applying MALF alone and our groupwise segmentation method (MALF+groupwise). The results are shown in terms of average number of mislabeled voxels produced for each hippocampus. Fig. 2 shows some results produced by the two methods. As expected, the performance of MALF increases as the number of atlases increases. Our groupwise method produced consistently better results than applying MALF alone. The error reduction rates caused by groupwise segmentation vary from ∼10% to ∼30%.

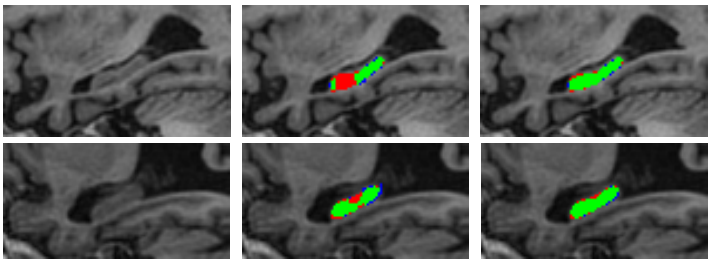

Fig. 2. Sagittal views of hippocampus segmentation. Red: manual segmentation; Blue: automatic segmentation; Green: overlap between manual and automatic segmentation

Fig. 3 (left) shows the Dice similarity coefficient (DSC)($\frac{2|A\cap B|}{|A|+|B|}$) for controls, patients with MCI and AD, respectively. When only one or two atlases were used by MALF to produce initial segmentations, the improvement by our groupwise method is about 10 % DSC. Our results using one atlas are better than

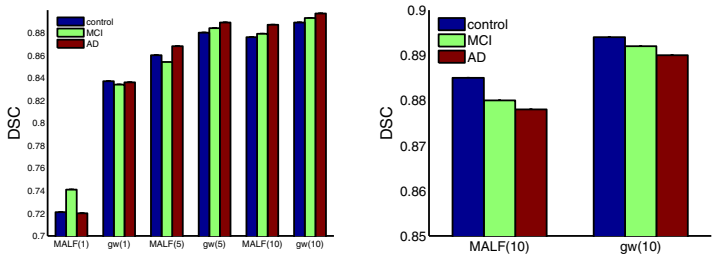

Fig. 3. Segmentation accuracy produced by using randomly selected atlases (left) and normal control atlases (right). The number of atlases used by MALF and our groupwise method (gw) is given in parentheses. Results are averaged over 5 cross validations.

the hippocampus segmentation results, ∼0.76, reported in [12], which also performed groupwise segmentation using one training image. When 5 and 10 atlases were used, the improvements caused by our groupwise approach are >2% and >1% DSC, respectively. To further test the generalization performance, we also repeated 5 cross-validation experiments, each with randomly selected 10 normal controls as atlases and 50 randomly selected subjects as testing images. As shown in Fig. 3 (right), our groupwise segmentation method produced an average DSC of 0.892, 1 % improvement over applying MALF alone. All improvements are statistically significant, with $p<0.01$ on the paired Students t-test for each cross-validation experiment. Our results also compare well to the state-of-the-art hippocampus segmentation performance, as summarized in Table 1[1]. Overall, our results compare favorably over the state-of-the-art but we used many fewer atlases than the competing works.

Table 1. Hippocampus segmentation performance in the recent literature

method : number of atlases used	Dice	JI	Tested Cohort
[13] : 38 atlases	0.87	-	normal control, AD
[4] : 79 atlases	0.887	-	normal control
[11]: 30 atlases	0.880	-	normal control
[8] : 55 atlases		0.80/0.81	normal control/MCI
[15] : 30 atlases	<0.85	-	normal control/MCI/AD
[14] : 20 atlases	0.892	-	normal control/MCI
[9] : 17 atlases	0.870	0.771	-
Our method : 10 atlases	0.893	0.805	normal control/MCI/AD

[1] Due to the differences in the images and manual segmentations used in different studies, quantitative comparisons across different publications may not be fair.

4 Conclusions and Discussion

We extended the powerful MALF technique to perform groupwise segmentation and validated our method in a hippocampus segmentation task. One drawback of groupwise segmentation is the additional computational cost for registrations among the testing images. However, this added cost is justified by the performance gain. For applications, where manually labeled atlases are limited and testing images are abundant, this technique will be more suitable to be applied.

References

1. Artaechevarria, X., Munoz-Barrutia, A., de Solorzano, C.O.: Combination strategies in multi-atlas image segmentation: Application to brain MR data. IEEE TMI 28(8), 1266–1277 (2009)
2. Bansal, R., Staib, L.H., Chen, Z., Rangarajan, A., Knisely, J.P.S., Nath, R., Duncan, J.S.: Entropy-based, multiple-portal-to-3D CT registration for prostate radiotherapy using iteratively estimated segmentation. In: Taylor, C., Colchester, A. (eds.) MICCAI 1999. LNCS, vol. 1679, pp. 567–578. Springer, Heidelberg (1999)
3. Besag, J.: Statistical analysis of non-lattice data. J. R. Statist. Soc. B 24(3), 179–195 (1975)
4. Collins, D., Pruessner, J.: Towards accurate, automatic segmentation of the hippocampus and amygdala from MRI by augmenting ANIMAL with a template library and label fusion. Neuroimage 52(4), 1355–1366 (2010)
5. Heckemann, R., Hajnal, J., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. Neuroimage 33, 115–126 (2006)
6. Jia, H., Yap, P., Shen, D.: Iterative multi-atlas-based multi-image segmentation with tree-based registration. Neuroimage 59(1), 422–430 (2012)
7. Kapur, T., Yezzi, L., Zollei, L.: A variational framework for joint segmentation and registration. In: IEEE CVPR - MMBIA, pp. 44–51 (2001)
8. Leung, K., Barnes, J., Ridgway, G., Bartlett, J., Clarkson, M., Macdonald, K., Schuff, N., Fox, N., Ourselin, S.: Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's Disease. Neuroimage 51, 1345–1359 (2010)
9. van der Lijn, F., de Bruijne, M., Klein, S., den Heijer, T., Hoogendam, Y.Y., van der Lugt, A., Breteler, M.M., Niessen, W.J.: Automated brain structure segmentation based on atlas registration and appearance models. IEEE Transactions on Medical Imaging 31(2), 276–286 (2012)
10. Lord, N.A., Ho, J., Vemuri, B.C.: Ussr: A unified framework for simultaneous smoothing, segmentation, and registration of multiple images. In: ICCV (2007)
11. Lotjonen, J., Wolz, R., Koikkalainen, J., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. Neuroimage 49(3), 2352–2365 (2010)
12. Raviv, T.R., Leemput, K.V., Menze, B., Wells, W.M., Golland, P.: Joint segmentation of image ensembles via latent atlases. MedIA 14, 654–665 (2010)
13. Sabuncu, M., Yeo, B., Leemput, K.V., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. IEEE TMI 29(10), 1714–1720 (2010)
14. Wang, H., Suh, J.W., Das, S., Pluta, J., Craige, C., Yushkevich, P.: Multi-atlas segmentation with joint label fusion. IEEE Trans. on PAMI 35(3), 611–623 (2013)
15. Wolz, R., Aljabar, P., Hajnal, J., Hammers, A., Rueckert, D.: Leap: Learning embeddings for atlas propagation. Neuroimage 49(2), 1316–1325 (2010)